基于可变粒度机会调度的网络大数据知识扩充算法 *

黄金国 [1]，刘　涛 [1]，周先春 [2]，严锡君 [3]

(1. 江苏开放大学 信息与机电工程学院，南京 210017; 2. 南京信息工程大学 电子与信息工程学院，南京 210017; 3. 河海大学 计算机与信息学院，南京 210098)

摘　要：为了满足网络大数据背景下，大数据传播的数据知识高精度要求和清除劣质数据干扰，基于粒度可变调整方案提出了机会调度的网络大数据知识扩充算法。在分析网络大数据特征基础上，通过自适应向量编码，捕捉网络大数据的异构特性，采用多阶反向传播将异构网络大数据归一化处理，再通过机会调度实现网络大数据实时传输。同时，基于网络大数据组成的知识工程系统分割细粒度大数据，将多维特征进行降维处理，使得知识粒度转变为已知，接着调整粒度动态特性，使得知识工程的大数据集具有线性特征和明确的几何特性，通过知识扩充提高知识获取精度。实验结果通过与基于细粒度的知识获取算法进行对比，证明了所提算法的网络数据传输的高可靠性、实时性和知识获取的高效率。

关键词：网络大数据；知识工程；知识扩充；可变粒度；机会调度

中图分类号：TP393　　doi: 10.3969/j.issn.1001-3695.2017.09.0947

Network big data knowledge extension algorithm based on variable granularity and opportunistic scheduling

Huang Jinguo[1], Liu Tao[1], Zhou Xianchun[2], Yan Xijun[3]

(1. *School of information &mechanical and electrical engineering Jiangsu Open University Nanjing* 210017 *China*; 2. *School of electronic &information engineering*, *Nanjing University of Information Science & Technology*, *Nanjing* 210017 *China*; 3. *College of computer & information Hohai University*, *Nanjing* 210098, *China*)

Abstract: In order to meet the needs of the network under the background of big data, and eliminate inferior data interference data knowledge high precision requirements of large data transmission, variable size adjustment scheme based on the algorithm to expand the network of large data knowledge opportunistic scheduling is proposed. Based on the analysis of large data network characteristics, the adaptive vector encoding, capture the heterogeneous characteristics of large data network, using multi order back-propagation network of heterogeneous data is normalized, and then through the real-time transmission of large data network to achieve opportunistic scheduling. At the same time, the knowledge engineering system composed of network data segmentation of fine-grained big data based on the multidimensional feature dimension, the granularity of knowledge transformation is known, then adjust the size of the dynamic characteristics, making big data set of knowledge engineering with linear characteristics and clear geometric characteristics, improve the accuracy of knowledge acquisition through knowledge expansion. The experimental results are compared with the algorithm based on fine grained knowledge acquisition, which proves the high reliability, real time and high efficiency of network data transmission.

Key Words: Network big data; knowledge engineering; knowledge extension; variable granularity; opportunistic scheduling

0　引言

网络大数据固有的类型异构、数据多元和分布式传播等特点[1,2]，使得如何在网络大数据背景下，确保大数据传播的数据知识精度[3]和消除劣质数据[4]干扰，获取有效解决问题的大数据知识结果成为知识工程[5]的关键问题之一。如何从网络大数据传播[6]的知识中重建数据知识库，得到了广泛关注。

网络大数据传播方向，文献[7]所提出的一种空间结构方案通过将符号型据被转换为值型，使得不仅有效保持原符号型特征而且重造了样本的相似度。文献[8]的优化原型系统，一方面

基金项目：国家自然科学基金资助项目（11202106，61201444）；江苏省高校自然科学研究面上基金资助项目（15KJD520003）

作者简介：黄金国（1976-），男，江苏泰兴人，副教授，硕士，主要研究方向为数据挖掘技术（huangjinguols@163.com）；刘涛（1980-），男，安徽芜湖人，副教授，硕士，主要研究方向为数据挖掘、无线传感网络等；周先春（1974-），男，安徽庐江人，副教授，博士，主要研究方向为信号与信息处理；严锡君（1963-），男，江苏南京人，副教授，博士，主要研究方向为数据挖掘与无线传感器网络.

可以加速多批量数据传输服务器集群，另一方面能够最好地利用带宽和分散随机线性网络编码的最大有用的信息传播。文献[9]研究了资源受限的移动机会网络的最优数据分发问题，解决了移动机会网络中时延受限的最小代价组播问题。

网络调度方向，基于随机线性网络编码，文献[10]提出了一种优先级调度方案，不仅可以利用信息包接收状态等线性关系反馈信息，还可以求解中继节点的有效信息规模。文献[11]研究了多通道无线链路之间的节点及其调度方案。文献[12]提出了一个完全分散的新分布式调度策略，使得每个节点根据其流量需求确定要调度的单元数量。

知识工程方向，文献[13]通过研究虚拟地理环境的地理知识特点，研究了虚拟地理环境的地理知识的分类及其工程架构。文献[14]提出了跨学科和多文化的方法来应对知识社会中的问题和挑战。文献[15]研究了一个知识工程框架处理零散的知识建模和多信息源的在线学习，对零散知识的非线性融合，和自动化需求驱动的知识导航。

在上述网络大数据传播和知识挖掘等领域的研究基础上，结合机会调度的网络大数据模型，研究了一种可以提高网络大数据传播效率和数据质量的知识扩充算法。

1 机会调度的网络大数据模型

与传统的网络数据相比，网络大数据具有一些明显特征例如数据类型复杂且异构、数据结构差异性、数据挖掘复杂度高和网络调度难度大等。针对上述网络大数据的特征，通过自适应向量编码，捕捉网络大数据的异构特性和类型特征，采用多阶反向传播统一异构网络大数据，通过机会调度网络大数据。

首先，将网络大数据挖掘对象从一维向量组转变为多维向量编码空间，得到网络大数据异构特征驱动的自适应多维向量编码模型。其次，基于大数据规模和维度，激励网络大数据在多维向量编码空间的异构特性和类型特征的捕捉。接着，提出了基于网络大数据挖掘对象和机会调度的多阶反向传播算法，将多维向量编码空间与多阶反向传播空间有机融合。最后，在多维向量编码空间中进行特征捕捉与多阶反向传播，从而组建了机会调度的网络大数据模型。

设一个 m 维有限域欧式空间 G^m，任意一维的向量空间为 A_i，$a_i^1 \in A_i^m, a_i^2 \in A_i^m, \cdots, a_i^m \in A_i^m$。定义 $A_i^1, A_i^2, \cdots, A_i^m$ 的多维向量编码空间的定义如下：

$$\begin{bmatrix} a_1^1 & \cdots & a_i^1 \\ \vdots & \ddots & \vdots \\ a_1^m & \cdots & a_i^m \end{bmatrix} \begin{bmatrix} A_i^1 \\ \vdots \\ A_i^m \end{bmatrix} = \begin{bmatrix} A_i^1 \sum_{j=1}^{i} a_j^1 \\ \vdots \\ A_i^m \sum_{j=1}^{i} a_j^m \end{bmatrix} \tag{1}$$

在 G^m 空间上，针对 A_i^m 的多维向量编码空间，网络大数据异构特征向量 B 与多维向量的映射关系如下：

$$\begin{cases} b_i^j = \sum_{i=1, j=1}^{m} \left\| A_i^m \right\|^2 \left(a_j^i - m \right) \\ a_i^j = \left\| B \right\|^2 + \sum_{i=1}^{m} A_i \left\| b_i \right\| \\ c = A \cdot B = \sum_{i=1, j=1}^{m} b_i^j \cdot a_i^j \cdot \left(\left\| A_i^m \right\|^2 + \left\| B \right\|^2 \right) \end{cases} \tag{2}$$

其中:b 表示向量 B 的元素，j 表示空间上的向量维度，c 表示基于向量 B 驱动，对向量 A 进行向量编码后的向量。

自适应多维向量编码形式描述如式（3）：

$$\begin{cases} C_{a_1 \cdots a_m} = \left\| A^m{}_{1,\cdots,m} \right\| + \prod_{i=1,\cdots,m} B_i \\ c_i^j = \dfrac{\delta}{2\pi} \exp\left\{ - \left\| a_i^j - b_i^j \right\|^2 \right\} \\ d_i^j = \dfrac{\sqrt{\left(a_i^j - b_i^j \right)}}{\left\| C \right\|^2} \end{cases} \tag{3}$$

其中:向量 $C_{a_1 \cdots a_m}$ 为多维向量空间上网络大数据的编码。变量 δ 表示向量维度偏移量。参量 d_i^j 表示向量 A 与向量 B 多维空间编码后之间的差异。

自适应多维向量编码后的网络大数据在网络传播过程中，进行如式（4）所示的迭代计算。

$$\begin{cases} \overline{C_{a_1 \cdots a_m}} = C_{a_1 \cdots a_m} - \dfrac{\int c_i \cdot \left\| A^m{}_i \right\|_{i=1,\ldots,m} dc}{\prod_{i=1,\cdots,m} B_i} \\ \overline{c_i^j} = c_i^j - \dfrac{\sum_{i=1, j=i}^{m} \left\| a_i^j - b_i^j \right\|^2}{2\pi} \end{cases} \tag{4}$$

其中: $\overline{C_{a_1 \cdots a_m}}$ 和 $\overline{c_i^j}$ 表示多维空间编码向量及其参量的网络传播的迭代变形处理。

为了提高网络大数据的传播效率，采用机会调度，机会参量可由式（5）计算得到。

$$o_{i=1\ldots m} = \frac{1}{m} \sum_{i=1}^{m} \frac{\int t \cdot \sqrt{\left| a_i^j - b_i^j \right|} dt}{\left\| C_{a_1, \ldots a_m} \right\|^2} \tag{5}$$

其中: $o_{i=1\ldots m}$ 表示 m 个维度上每个维度的机会调度权重。参数 t 表示网络大数据传播时间。随着网络大数据编码传播的过程，基于多维空间编码差异进行网络大数据传播的机会调度。

综上所述，网络大数据的多阶反向传播的主要步骤如下：

a)获取网络大数据的多维空间编码及其参量，得到 $C_{a_1 \cdots a_m}$ 和 c_i^j 。

b)对多维空间编码向量及其参量的进行网络传播的迭代变形处理，得到 $\overline{C_{a_1 \cdots a_m}}$ 和 $\overline{c_i^j}$ 。

c)对于每一维空间的编码参量 c，求解机会调度权重 $o_{i=1\dots m}$。

d)对于每一阶网络大数据传播，求解前向集合 F_i^m 和反向数据集合 R_i^m。

e)计算向量 A 与向量 B 的多维空间编码后之间的差异 d_i^j。

f)计算 F_i^m 和 R_i^m 的残差，修正反向网络大数据集合 $\overline{R_i^m}$。

机会调度的网络大数据的多阶反向传播算法描述如下：

输入：m，A_i

输出：$\overline{R_i^m}$

for i=1, i++, i<=m

$$\left\|A^m{}_{1,\cdots,m}\right\|;$$

for j=1, j++, j<=m

$B_j=B_j*B_{j-1}$;

obtain $C_{a_1\cdots a_m}$ and c_i^j;

for i=1, i++, i<=m

$$\overline{C_{a_1\cdots a_m}}=C_{a_1\cdots a_m}-\frac{\int c_i\cdot\left\|A^m{}_i\right\|_{i=1,\dots,m}dc}{\prod_{i=1,\cdots,m}B_i};$$

$$\overline{c_i^j}=c_i^j-\frac{\sum_{i=1,j=i}^{m}\left\|a_i^j-b_i^j\right\|^2}{2\pi};$$

for i=1, i++, i<=m

temp= $\sum_{i=1}^{m}\frac{\int t\cdot\sqrt{\left|a_i^j-b_i^j\right|}dt}{\left\|C_{a_1,\dots a_m}\right\|^2}$;

$o_{i=1\dots m}$=temp/m;

computing the F_i^m and R_i^m;

amending the R_i^m with F_i^m;

return $\overline{R_i^m}$;

2 可变粒度的知识扩充算法

基于网络大数据组成的知识工程系统定义为三元组 K=<R, E, AR>，其中 R 表示网络大数据集；E 表知识描述对象；AR 表示大数据集所有元素的知识属性集。对于，$\forall ar\in AR$，$\forall e\in E$，定义线性关系属性映射 $ar:e\rightarrow T_r$，其中 T_r 表示网络大数据集 R 的任一元素 r 的知识属性映射关系，形如 $R(e)\subset AR(r)$。因此，一个粗糙的知识工程系统可定义为 $KR=\left\langle R,E\cap AR(r),AR\cup\lambda\right\rangle$，其中 λ 表示粒度粗糙权重。

假设 K 是一个多粒度粗糙知识工程系统，R，E 和 AR 之间存在模糊粗糙映射关系，且 R 与 E 之间的元素映射存在多对多现象，即 $h\in R\cap E$ 且 $h\in\forall ar\subset AR$。于是，K 中的细粒度精确知识集合 $\overline{K}$ 与粗糙知识工程系统 KR 的细粒度知识集合 $\overline{KR}\left(\lambda<\overline{\lambda}\right)$，存在如式（6）所示的关系。

$$\begin{cases}\overline{K}=\left\{r\in R:\frac{(e,ar)\cup(r\rightarrow AR(r))}{e}\subset K\right\}\\ \overline{KR}=\left\{r\in R:\frac{1}{r}(e,ar)\cap(r,\lambda)\cap K\right\}\\ \overline{\lambda}=\frac{\sum_{i=1}^{m}\sqrt{\left|a_i-b_{ii}\right|}}{\left\|R\right\|^2}\end{cases}\tag{6}$$

其中: $\overline{\lambda}$ 表示粗糙集的细粒度阈值。

在网络大数据知识工程系统中，知识粒度在大数据网络传输的多维向量空间中具有多维特征，该特征使得知识粒度具有未知性和动态特性。为了寻求细粒度大数据，以便将多维特征进行降维处理，使得知识粒度转变为已知，且调整动态特性，使得知识工程的大数据集具有线性特征和明确的几何特性，提高知识挖掘精度和知识处理目标的唯一定义。对于，粗粒度多维大数据，通过调整粒度特征和降维处理，降未知性进行线性描述，隐藏未知大数据的多维空间几何特性。多维向量空间与细粒度知识几何特征空间之间的对应关系详见图 1，知识工程的三元组降维至二元组，将未知因素进行了确定性转换和几何特征。

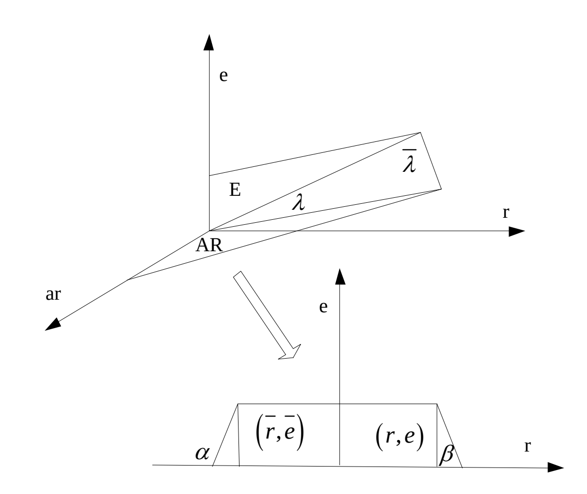


图 1 多维空间与几何特征空间的对应

因此，对于知识工程系统 K，基于可变粒度给出网络大数据知识的参数和属性描述：

$$\begin{cases}\overline{r}=f(r,E,AR),r\in\overline{R_i^m},f:R^m\rightarrow V^m\\ f(r,E,AR)=\frac{(E,AR)\cap(r,\lambda)}{r\left\|K\right\|^2}\end{cases}\tag{7}$$

其中，$f:R^m\rightarrow V^m$ 表示从原始多维向量空间 R^m 到可变力度特征空间 V^m 的降维映射。

可变粒度转换可以通过方程 ρ=(r*sinα+e*cosβ)f(r, e,ar)进行解析完成，其中 ρ 表示可变粒度，α 表示任一大数据大数据元素 r 的多维向量空间水平交叉弧度，β 表示任一知识描述对象 e 在空间降维过程中产生的垂直交叉弧度。因此，可变粒度与网络大数据知识工程的迭代关系如式（8）所示。

$$\begin{cases} (r,e) \mid \rho = f(r,e;\alpha) & \{(r,e)\in R^m\} \\ (\bar{r},e) \mid \rho = f(\bar{r},e;\beta) & \dfrac{\rho}{\lambda} \le \cos\beta \end{cases} \tag{8}$$

降维后知识平面上的数据点经过粒度可变转换后，多维空间的知识集全部转入细粒度几何特征空间。该空间内的数据知识具备了确定关系和线性特征。此时，网络大数据知识工程系统 KR 有效解决了粗粒度的不规则几何空间对知识挖掘的干扰和粒度的动态变化对知识空间的影响。图 2 给出了网络大数据知识获取中的粒度调度方案，以 $\bar{\lambda}$ 为阈值分割粗粒度集和细粒度集。细粒度直接进入获取结果，粗粒度通过可变粒度调度，消除未知性和不规则性，转换为细粒度。

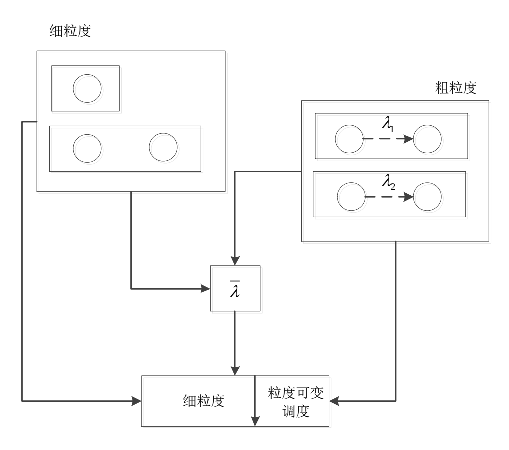


图 2　基于可变粒度的知识获取模型

经过图 2 的可变粒度调度后，网络大数据知识工程在 $\bar{\lambda}$ 的细粒度分割后，通过式（9）进行知识扩充。

$$\begin{cases} f\left(\overline{R_i^m}, \bar{\lambda}, \rho, \alpha, \beta\right) = \dfrac{\sum_{i=1}^{m} r_i + \bar{\lambda}(\sin\alpha + \cos\beta)^{\rho}}{\|R\|^2} \\ \alpha = \arctan \dfrac{\lambda}{\bar{\lambda}} \\ \beta = \arctan \sqrt{\left|\rho^2 - \bar{\lambda}\right|} \end{cases} \tag{9}$$

3　实验结果分析

对基于可变粒度机会调度的网络大数据知识扩充算法记为 NKE-VOS 进行性能分析与验证。实验中，主要分析了网络调度后数据误差、数据传输延迟、知识获取的收敛次数等性能。在相同实验环境下，所提算法的上述性能与基于细粒度的知识获取算法记为 FGKA 进行了对比。所采用的实验平台如表 1 所述。

表 1　实验平台

参数	取值
网络终端数	50 个
网络服务器数	5 个
服务器 CPU	Intel Xeon E3 v2
服务器硬盘空间	2 TB
无线通信协议	IEEE 802.11g
服务器操作系统	Ubuntu Server 16.04.2 LTS
算法开发语言	Java
实验时间	50 min
网络终端存储空间	4 GB

图 3 给出了逐步激活网络终端后，随着网络大数据量的增加，两种算法所采用的网络调度算法在数据精度方面的表现。两种算法在 50 min 内调度传输的大数据分别与原数据进行比对，得到数据误差。对比发现，FGKA 算法的所采用的静态调度，对与大数据的规模变化反映迟缓，导致数据丢失或出错，严重制约了数据质量。反之，所设计的 NKE-VOS 算法所采用的机会调度，将网络大数据挖掘对象从一维向量组转变为多维向量编码空间，基于大数据规模和维度，激励网络大数据在多维向量编码空间的异构特性和类型特征的捕捉，实现高效率网络调度，有助于提高数据精度。

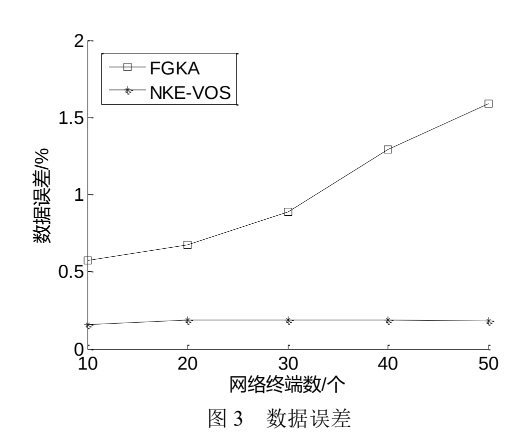


图 3　数据误差

图 4 给出了随着网络大数据量的增加，两种算法在网络传输实时性方面的表现。分别统计了 50 分钟内两种算法传输的大数据的端到端延迟，并求平均值。对比发现，所提出的 NKE-VOS 算法通过获取网络大数据的多维空间编码及其参量，接着对每一维空间的编码参量进行机会调度，对于每一阶网络大数据传播，求解前向集合和反向数据集合，采用机会调度的网络大数据的多阶反向传播算法，从而缩短网络数据传输延迟，保障实时性。

图 5 给出了随着网络服务器的增加，两种算法完成知识获取所需要的迭代次数。所提 NKE-VOS 算法将粗粒度多维大数据进行粒度可变调度，同时降维，明确未知性的线性描述，消除未知大数据，重构多维空间几何特性，所以可以在较少的迭代过程中获取知识，从而扩充网络大数据知识。

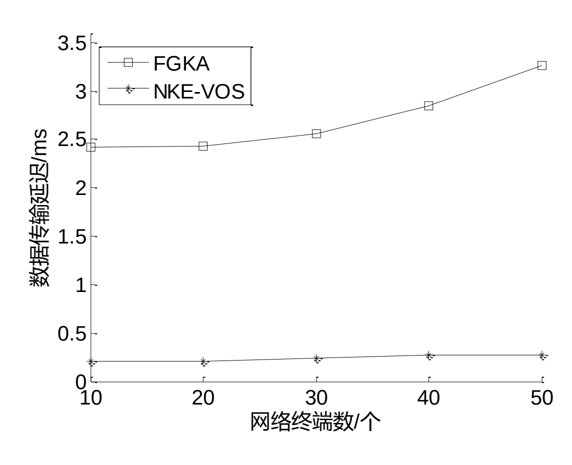


图 4　数据传输延迟

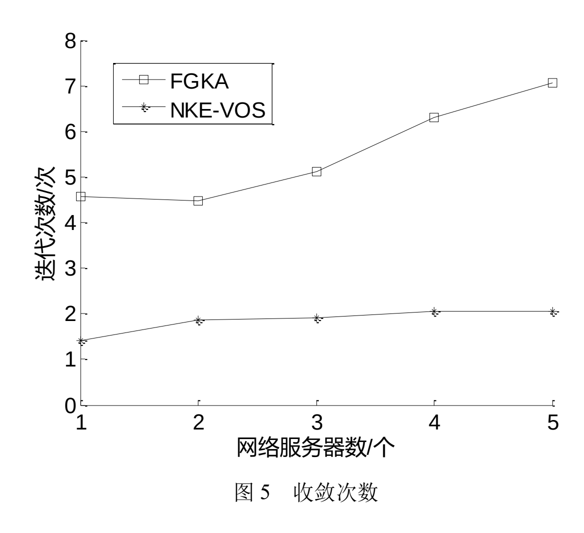


图 5　收敛次数

4　结束语

网络大数据应用对数据传播的实时性、数据精度和知识获取效率提出了更高要求，为了满足上述要求，提出了一种适用于网络大数据的基于可变粒度和机会调度的网络大数据知识扩充算法。首先，该算法从网络大数据异构特征出发，建立多维向量空间，实时捕捉异构特性，通过自适应多维向量编码，经过多阶反向传播和机会调度，保障网络大数据传输的实时性和可靠性。其次，将网络大数据的知识工程系统按照粒度可变阈值进行细粒度分割，通过多维特征的降维，实现知识粒度的已知性和动态几何特征的明确，采用了基于可变粒度的知识扩充算法。所提算法与基于细粒度的知识获取算法在数据误差、数据传输延迟、知识获取的收敛次数等性能的对比实验，结果表明所提算法在网络大数据传输可靠性、实时性和知识获取效率等方面具有明显优势。

参考文献：

[1] Gao J, Lu W F, Dai Z J, et al. A computational approach to characterizing the impact of social influence on individuals'vaccination decision making. [J]. Plos One, 2013, 8 (4): 601-611.

[2] Przulj N, Malod-Dognin N. Network analytics in the age of big data [J]. Science, 2016, 353 (6295): 123-124.

[3] Gao S, Pang H, Gallinari P, et al. A novel embedding method for information diffusion prediction in social network big data [J]. IEEE Trans on Industrial Informatics, 2017, 3 (99): 1-11.

[4] Kim R, Lim H, Krishnamachari B. Prefetching-based data dissemination in vehicular cloud systems [J]. IEEE Trans on Vehicular Technology, 2016, 65 (1): 292-306.

[5] 康文文，李浩敏，汤超，等. 面向复杂工程系统设计的云知识平台 [J]. 系统工程与电子技术, 2017, 39 (5): 1078-1084.

[6] Zheng X, Wang J, Dong W, et al. Bulk Data dissemination in wireless sensor networks: analysis, implications and improvement [J]. IEEE Trans on Computers, 2016, 65 (5): 1428-1439.

[7] 王齐，钱宇华，李飞江. 基于空间结构的符号数据仿射传播算法 [J]. 模式识别与人工智能, 2016, 29 (12): 1132-1139.

[8] Liu Y, Niu D, Khabbazian M. Cooper: Expedite Batch Data Dissemination in Computer Clusters with Coded Gossips [J]. IEEE Transon Parallel & Distributed Systems, 2017, 28 (8): 2204-2217.

[9] Liu Y, Wu H, Xia Y, et al. Optimal Online Data Dissemination for Resource Constrained Mobile Opportunistic Networks [J]. IEEE Trans on Vehicular Technology, 2017, 66 (6): 5301-5315.

[10] 王练，梁申虎，彭代渊. 多源多中继无线网络中基于随机线性网络编码的调度方案 [J]. 电子与信息学报, 2017, 39 (3): 532-538.

[11] Moharir S, Krishnasamy S, Shakkottai S. Scheduling in densified networks: algorithms and performance [J]. IEEE//ACM Trans on Networking. 2017, 25 (1): 164-178.

[12] Domingo-Prieto M, Chang T, Vilajosana X, et al. Distributed PID-based scheduling for 6tisch networks [J]. IEEE Communications Letters, 2016, 20 (5): 1006-1009.

[13] 林珲，游兰. 虚拟地理环境知识工程初探 [J]. 地球信息科学学报, 2015, 17 (12): 1423-1430.

[14] García-Peñalvo F J. Engineering contributions to a multicultural perspective of the knowledge society [J]. IEEE Revista Iberoamericana De Tecnologias Del Aprendizaje, 2015, 10 (1): 17-18.

[15] Wu X, Chen H, Zhang Q, et al. Knowledge Engineering with Big Data [J]. IEEE Intelligent Systems, 2015, 30 (5): 46-55.